DOCUMENT RESUME

ED 199 995 FL 011 837

AUTHOR Rand, Earl
TITLE The Effects of Test Length and Scoring Method on the Precision of Cloze Test Scores.
INSTITUTION California Univ., Los Angeles. Dept. of English.
PUB DATE Jun 78
NOTE 11p.; In its Workpapers in Teaching English as a Second Language, Volume XII, p62-71, Jun 1978.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cloze Procedure; English (Second Language); Higher Education; Language Research; *Language Tests; Multiple Choice Tests; *Scores; *Scoring Formulas; *Test Validity

ABSTRACT

A project is described that was undertaken to investigate: (1) how long a cloze test has to be to achieve optimum reliability without wasting anyone's time; and (2) how cloze tests should be scored in order to obtain maximum reliability. The literature recommended 50 deletions in order to provide for an adequate sample of examinee's abilities; it did not strongly recommend either exact-word scoring or acceptable-word scoring. A 50-deletion regular cloze test and a multiple-choice cloze test were administered, respectively, to 50 and 57 foreign students. The regular test answer sheets were scored by the exact-word, acceptable-word, and clozentropy methods. It is concluded from examination of the data that the acceptable-word method is superior to the other two methods and to the multiple-choice cloze test. It was also shown that by 25 items, the maximum reliability had almost been achieved; at the same time it was clear that the best ratio is obtained through tests with 50 items. It is recommended that the most efficient use of everyone's resources can be achieved in a cloze test of 25 deletions and use of the acceptable-word method of scoring. (AMH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Workpapers in Teaching English as a Second Language, Vol. XII

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Donald Bowen, Acting
Vice-Chairman, Dept of Eng
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED199995

FL011837

THE EFFECTS OF TEST LENGTH AND SCORING METHOD ON THE PRECISION OF CLOZE TEST SCORES

Earl Rand

Precision in tests and measurements refers to the tendency of an examinee's observed scores to cluster closely about his true score.[1] Thus precision is related inversely to the variance of the distribution of scores, i.e., the smaller the variance, the greater the precision. Figures 1 and 2 illustrate this concept by plotting the scores on two sixty-item tests taken a number of times by the same person. Test A has more precision than Test B because the examinee's observed scores are more closely clustered.

Variance[2] of test scores results from a number of factors. Variance increases (1) as a test is lengthened, (2) as the mean difficulty of each item approaches .5, and (3) as the intercorrelation of items increases. Thus long tests, with highly intercorrelating items which half the examinees get right, have maximum dispersion of scores. The above three factors also increase reliability (consistency and homogeneity). Thus, for maximum reliability, a test should be long, with highly intercorrelating items. And, again, the average item difficulty should be close to .5.

Two questions are answered here: first, how long should a cloze test be to achieve optimum reliability without wasting examinee's time and effort? And second, how should cloze tests be scored in order to obtain maximum reliability?

The purpose of this paper is twofold. It argues first that length of a cloze test need not exceed 25 deletions and second that acceptable-word scoring is superior to exact-word scoring. (As a side issue, it supports Jonz's research (1976) showing that a multiple choice cloze test can provide reliable measures of ability.)

Length

Most studies either utilizing or investigating cloze tests contain fifty deletions. Although fifty deletions seems to be standard, it hasn't been shown that fifty blanks is the minimum for adequately measuring the examinee's ability and yet optimally utilizing his effort and the testing time. For example, Haskell states that with passages ranging from thirty-five to seventy-five deletions, "The number of deletions within this range did not effect the test results" (p. 2-3). Bormuth (1964) claims that

cloze tests with fewer than fifty items tend to to be unreliable because the means on different cloze test forms made from the same passage differ significantly. (Bormuth's "unreliable" refers to the confidence one can place in the equivalence of group means. (See Amal Mahmoud's research, cited below.) In classical test theory, reliability refers to the consistency of individual rankings and/or homogeneity of test items, i.e., to what extent are all the items measuring the same underlying ability?

Cloze tests have been used as part of the UCLA ESL Placement Examination for the past seven or eight years. They have had KR-20 reliabilities from the mid seventies to the mid nineties (.73 to .94). And usually they have contained fifty deletions, with an average difficulty between .4 and .6, depending on reading level, scoring method, and amount of time allowed.[3]

Numerous studies, at UCLA and elsewhere, have shown that cloze tests provide useful scores for placing students in general levels or ESL programs. In fact, from these studies, one can conclude that just about any cloze test of adequate length and scored by any standard procedure will give much the same results as any other cloze test (with equivalent content[4] and at the same reading level). Also, the "same" test, with different deletions or scoring methods, correlates with "itself" very highly, although the mean scores may differ as much as twenty or thirty or more percentage points.[5]

Jongsma summarizes that "test length of 250 words with a minimum of 50 deletions has often been suggested, but some investigators have used much longer tests" (1970, p. 36). Amal Mahmoud in masters thesis directed by J. Donald Bowen (1977) recommends one hundred deletions. He found, like Bormuth (1964), that the means from equivalent tests differed substantially with fewer than 100 blanks. Both he and Bormuth are concerned with absolute equivalence of scores, whereas classical test theory is interested mainly in relative equivalence, i.e., two or more "parallel" tests should rank individuals in the same order, regardless of the absolute number of points made on each test. One can conclude that the literature recommends fifty deletions in order to provide for an adequate sample of the examinees' abilities, given that the passage content, reading level, and available time are appropriate.

Scoring Method

Besides length, method of scoring also directly influences variance and thus reliability. Jongsma (1970) begins his review of research on cloze scoring methods with the statement: "Probably the most widely researched aspect of cloze methodology has been that of scoring" (1971, p. 7).

He summarizes his section on cloze scoring research:

> Thus it appears that the literature consistantly shows that the scoring of exact replacements to be the most objective, efficient, and useful scoring system to use with the cloze test procedure. Although slightly higher reliability has been obtained at times using other procedures, such as synonym count, the increased time and subjectivity necessary for such systems do not warrant their use (1970 pp. 9-10).

Later he concludes, "Scoring exact replacements is the most efficient and useful procedure" (1970, p. 36). Bormuth (1965) concludes that scoring responses exactly matching the deleted word is the most valid scheme. Ruddell (1964), on the other hand, finds that exact-word and synonym scoring schemes do not differ. He reports reliabilities of .85 to .89 for exact-word scoring and .89 to .91 for acceptable alternative scoring. Hincfotis (1978) finds the reliabilities for her fifty-deletion test to differ greatly: exact-word = .61 and acceptable-word = .85. For practical reasons, she recommends exact-word scoring. Stubbs and Tucker (1974), not reporting reliabilities, find that the exact and acceptable-word methods correlate very highly but that the acceptable-word method correlates higher than the exact-word method with other measures of ESL proficiency, such as listening, structure, and vocabulary.

A third method of scoring is "clozentropy" (Darnell, 1970 and Reilly, 1971). Briefly, this modification of the acceptable alternative method requires that the examiner give the test to native speakers. The candidate receives points for an answer based on the natural log of the frequency of his choice. Thus, for example, if for a certain blank, seven native speakers chose 'machine', the score for answering 'machine' would be 1.94; if 19 chose 'washer', then the score would be 2.94. (The natural log of 7 is 1.94, and of 19, it is 2.94.) Without a computer, Darnell's clozentropy method involves a bit too many calculations for most cloze test users. The reliability of this method is reported to be as high as the other more common methods.

To conclude, the literature does not strongly recommend either exact-word or acceptable-word scoring. The examiner can, it seems, select either method of scoring and not worry about the results.

Method

James Brown (1978), as part of his MATESL thesis research, constructed a 50-deletion regular cloze test and a multiple-choice cloze test, which were administered to 55 and 57, respectively, foreign students as part of the UCLA ESL Placement Examination in January 1978, called "W-78". The regular test answer sheets were scored by the exact-word, acceptable-word, and clozentropy methods.

After the answer sheets were keypunched, Rand developed a special item analysis program in order to compute traditional item and test statistics for each method, while the test was being artificially lengthened by one item (deletion) at a time. Statistics for the M-C test were also computed this way. Thus, the computer output provided standard statistics for a two-item test, a three-item test, a four-item test, and so forth up to a fifty-item test for each scoring method.

Results

The reliability for each test length and each scoring method is presented in Figure 3. It is clear that the acceptable-word method is superior to the other two methods and to the M-C cloze test. It is also obvious that the reliability starts leveling off on most test scoring methods around twenty items. By twenty-five items, the maximum reliability has almost been achieved, even for the F-77 cloze scored by the exact-word method.

Figures 4 and 5 show the mean scores +/- one SEM and one SD. The band (confidence interval) made by the SEM for the acceptable word method is noticeably smaller, relative to the SD band, than for the exact-word method.

An examiner wants a narrow SEM and a wide SD. Thus, he can be confident that an examinee's true score is close to his observed score, and with a large SD, he can be confident that his test separates those examinees with high ability from those with low ability.

Figure 6 shows the ratio of SEM to SD. It clearly reveals that one obtains the very best ratio with tests containing fifty deletions. But it also is clear that the ratio levels off around twenty-five deletions. Little precision is gained by making a cloze test longer than twenty-five items.

Conclusions

This study has shown that the examiner using cloze tests can most efficiently use his own resources and his examinees' time by giving a cloze test of only twenty-five

deletions and employing the acceptable-word method of scoring.

Notes

[1]The user of test scores wants the obtained score to indicate precisely the "true" score. Of course, one never knows the true score, but with two statistics [the reliability and standard deviation (the square root of the variance)], he can compute the standard error of measurement, which allows him to compute the interval within which the true score probably occurs. This formula is

$$SEM = SD\sqrt{1 - r}$$

SEM= standard error of the mean
SD = standard deviation
r = reliability
The effect of r, ranging from 0 to +1, on SEM is obvious: the greater r is, the smaller SEM will be, relative to SD. The smaller r is, the nearer the portion under the radical sign will be to 1.0, and the larger SEM will be, relative to SD.

[2]Variance is the average squared difference between a set of scores and the average of the set. Variance is the mean of the squares of variation from the arithmetic mean. It is a statistic describing the dispersion of a set of scores.

[3]Multiple-choice subtests (e.g., Verb Forms, Romance Vocabulary, Listening) have usually been made up of 15 to 30 items, with a difficulty between .6 and .75, a bit on the easy side. Consequently, their KR-20's have been slightly lower. However, they give specific diagnostic information, are easy to grade, etc. (See Rand 1972 and 1973.)

[4] Ray Hoy (1975) has shown that content, however, interacts with the examinee's background.

[5]This high correlation would indicate that different deletions or scoring methods do not matter as far as relative score is concerned. But this is only partly true. One wants to know not only the examinee's score relative to other examinees' scores but also the amount of faith one can have that his "true" score is somewhere near. The SEM tells how near "somewhere" probably is.

References

Bormuth J. R. "Experimental applications of cloze tests," IRA Conf. Proc. 9(1964), 303-306.

Bormuth, J. R. "Validities of grammatical and semantic classification of cloze test scores," IRA Conf. Proc. 10(1965), 283-285.

Brown, James. "Correlational study of four methods for scoring cloze tests," UCLA MA Thesis, Unpublished, 1978.

Darnell, D. K. "Clozentropy: a procedure for testing English language proficiency of foreign students," Speech Monog. 37(1970), 36-46.

Haskell, John F. Putting cloze into the Classroom. Unpub. MSS. no date.

Hinofotis, Frances and B.G. Snow. "Cloze as an alternative method of ESL placement and proficiency testing," in J.W. Oller, Jr. Language in Education: Testing the Tests. Rowley: Newbury. 1978.

Jongsma, Eugene A. "The cloze procedure: A survey of the research." Indiana U., School of Education. Occasional papers in reading. n.d. c. 1970. ERIC ED050893

Jonz, Jon. "Improving on the basic egg: the M-C cloze," LL 26 (1976), 255-266.

Rahmoud, Amal "A study of the stability of alternate deletion patterns in cloze tests." American University in Cairo. MA Thesis, Unpublished, 1977.

Roy, Raymond H. The effect of vocabulary clues, content familiarity and English proficiency on cloze scores. UCLA. Unpub. MA thesis. 1975.

Oller, John "Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language,"

Potter, Thomas C. A taxonomy of cloze research, part 1: readability and reading comprehension. Inglewood, Calif.: SWRL Tech. Rep. #1. June 1968.

Rand, Earl. "Integrative and discrete-point tests at UCLA," UCLA TESL Workpapers. 6(1972), 67-78.

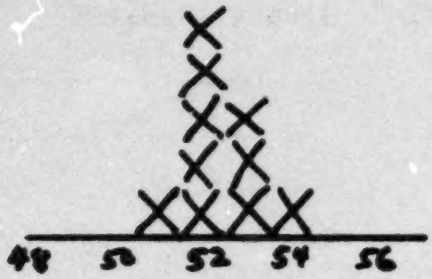
Rand, Earl. "Comments on 'a reply'," UCLA TESL Workpapers. 7(1973), 95-99.

Reilly, R.R. "A note on clozentropy," Speech Monographs.38(1971), 530-537.

Buddell, R. B. "A study of the cloze comprehension technique in relation to structurally controlled reading material," IRA Conf. Proc. 9(1964), 298-303.

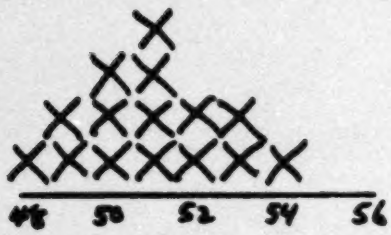
Stubbs, Joseph B. and G. Richard Tucker. "The cloze test as a measure of English proficiency," Modern Language Journal 58(1974), 239-241.

APPENDIX



TEST A

Figure 1



TEST B

Figure 2

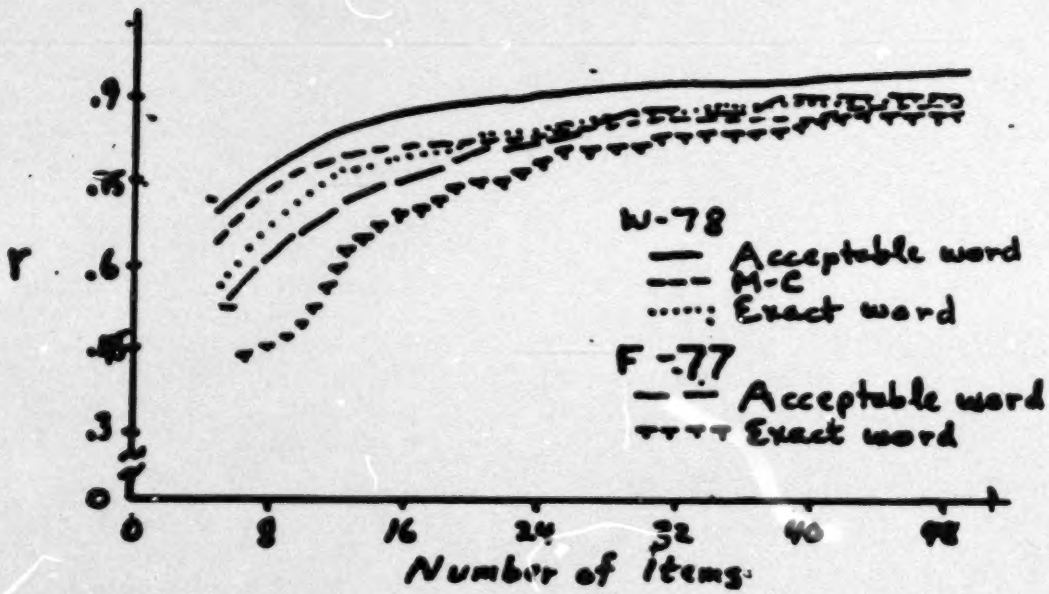


Figure 3

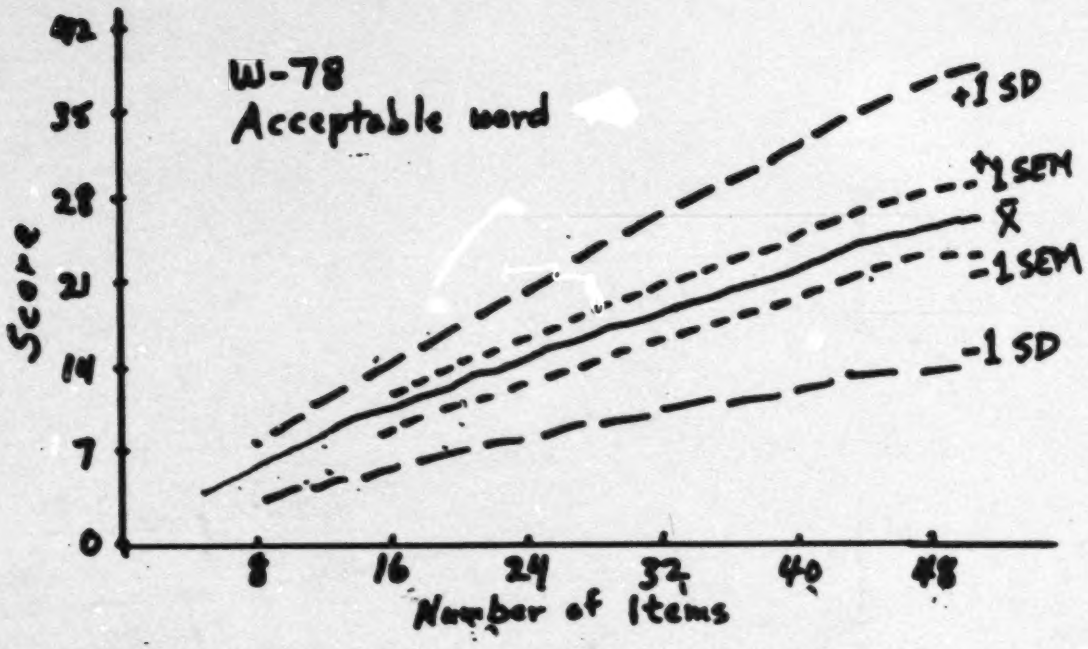


Figure 4

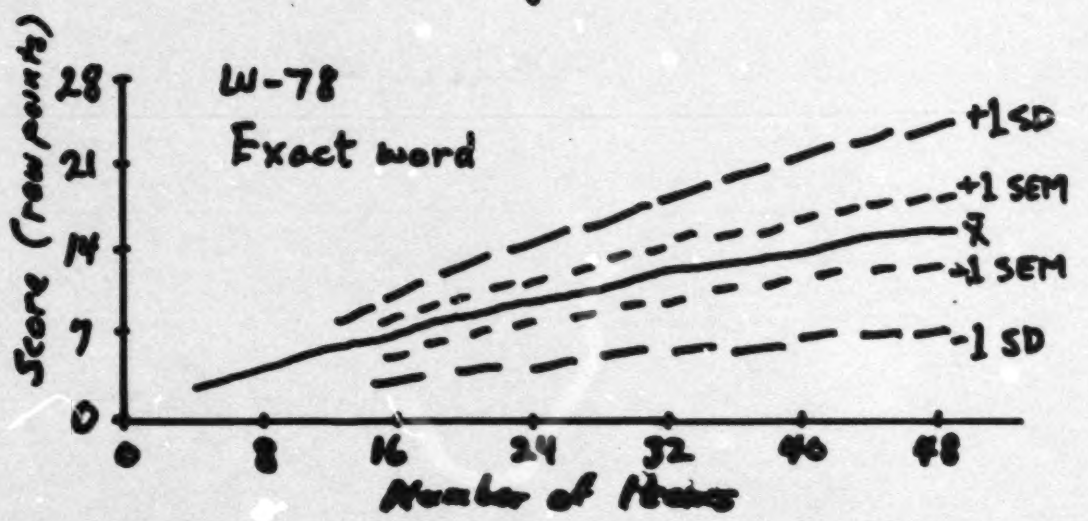


Figure 5

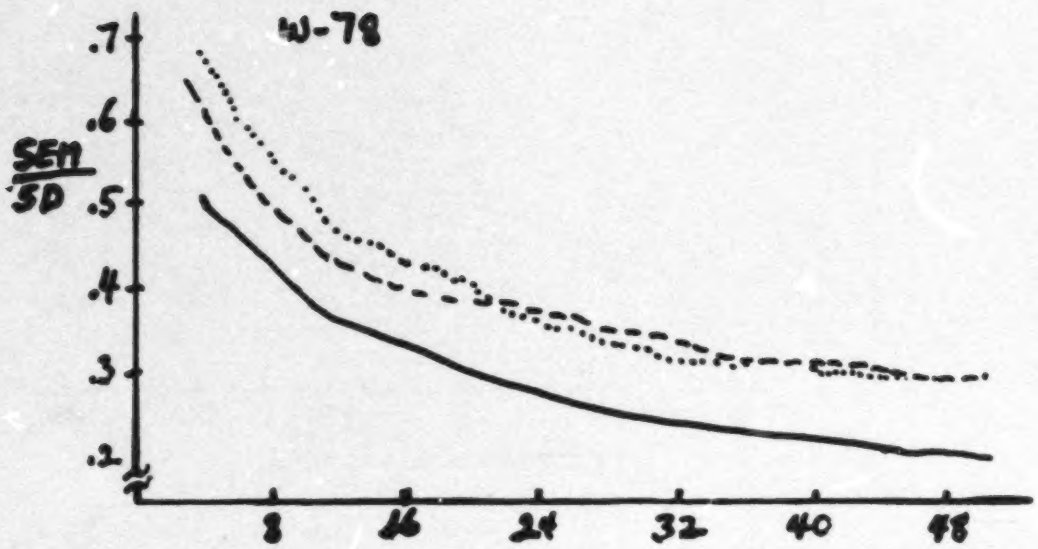


Figure 6